Data Wrangling Report on WeRateDogs Datasets

In order to create a worthy analyses and visulization, we wrangled datasets obtained from WeRateDogs, The wrangling process consists of three major steps which includes:

- Gathering Data
- Assessing Data
- Cleaning Data

The Data used in this project was gathered from various sources, including downloading manually for dataset that is already available, using the requests library to download additional dataset from the link provided and using tweepy API, to gathered additional data that was needed. This resulted in collecting three datasets, which were then loaded into three seperate pandas dataframe namely(WeRateDogs, api_tweets, and Image_prediction). These datasets sets were assessed for Quality and Tidiness issues using visual assessment using Excel and programmatic assessment using python methods such as; .info(), .head(), .describe(), .value_counts(), etc. The following problems were identified.

Quality issues

WeRateDogs Table

- rows containing retweeted comments and no images would be removed since their rating are not original.
- Columns such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp, retweeted_status_user_id, contains much missing values and also these columns would not be used in our analyses, hence they are not needed.
- Source Column in the WeRateDogs Table containing where the source of the tweet is in HTML format, this would be changed to clearly depict the source of the tweet.
- rows in the name column that contains dogs that do not have names are replaced with None, a, an, the, this would be changed to not named.
- The timestamp column is in string. This should be in datetime.
- Dogs whose dog stages were filled with None will be replaced with Not Specified. This is done to prevent any misunderstanding as None is not a dog stage.
- The value 10 in the rating denominator column accounts for 99% of the rating_denominator column data, so the value would be changed to 10, in order to create a new standard column for dog rating using the rating numerator and denominator column.
- Outliers were found in the rating numerator column, so for more accurate analyses, values above the upper bound of 15 would be removed and a new column named standard_rating would be created.

Tidiness issues

WeRateDogs Table

- Columns like doggo, pupper, puppo, and floofer are dog stages and should be in a single column.
- api_tweets containg retweet count and favourite count should be merged with WeRateDogs table which contains dog ranking, because retweet count and favorite count can be used to rank dogs. Also, the jpg_url column in Image_prediction table would be moved with the WeRateDogs table.

Tidiness issues were addressed first. The doggo, pupper, puppo and floofer columns in the WeRateDogs table were merged into one column since all these columns are dog stages. Atter this we dealt with the second Tidiness issue which was to merge the api_tweets table with the WeRateDogs Table, since the retweet count and favourite count can be used to rank dogs, also these columns form one observational unit with the WeRateDogs Table.

After dealing with the Tidiness issues, we proceeded to deal with quality issues. retweeted rows and rows that had no images were removed from our dataset since their ratings are retweeted and some contain negative ratings. Columns in the WeRateDogs Table that were not needed for our analysis were removed, these columns include; in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp, and retweeted_status_user_id column were removed. Also the source column in the WeRateDogs Tables was in HTML format, so the source of the tweet replaced the html source for each value. Rows in the name column of the WeRateDogs table that contained None, a, an and the which was used to represent dogs that were not named was changed to Not named, in order to ensure consistency. Also, dogs whose stages were filled with None was changed to Not specified. This was done to ensure clarity. Finally, since 99% of the values in the rating numerator was 10, all the values were changed to 10. Also all the outliers in the rating numerator column was changed to 15, with these changes made to the rating numerator and rating denominator columns, a new standard rating column was created.

In conclusion, The cleaned dataset was saved to a csv stored as twitter_archive_master.csv. Since data wrangling is an iterative process, some outliers were found and cleaned during the Analysing and visualization stage.